MLDS CENTER

Maryland Longitudinal
Data System

Better Data • Informed Choices • Improved Results

MLDS Synthetic Data Project:
An Evaluation

JSM 2019
July 29th, 2019

Outline

- The Maryland Longitudinal Data System (MLDSC)
- The Synthetic Data Project
- Synthetic data evaluation
  - Research utility
  - Disclosure risk

The MLDSC

- Receives, matches and merges education and workforce data from 3 partner state agencies: MSDE (grades 9-12), MHEC (postsecondary), & DLLR (wages)

- Mission is to produce research reports and dashboards to inform state policy, programming, and the public

- Confidential information
  - Data confidentiality protected by federal and state laws
  - Access granted to MLDSC staff only

The Synthetic Data Project

- In 2015, the State of Maryland received a grant from the U.S. Department of Education's State Longitudinal Data Systems program. A portion of these award dollars (about $2.7M) was to create a synthetic data system of the data in the MLDS.
    - Aim: Expand access to the data to leverage research value

- Synthetic data are generated based on models to mimic the relational patterns among variables, so statistical analyses with such "fake" synthetic data should yield findings substantially similar to the real data

- Simultaneously, reduces the risk of privacy breach

The Synthetic Data Project

- **Creation of a gold standard standard dataset (GSDS)**
  - **Study the data**
  - **Define GSDS**
- **Synthesize GSDS**
- **Evaluate research utility and disclosure risk** ←
- Governing Board Approval
  - Release
  - Allow users to send error free codes to be run on real data
- Report on the project to inform other state longitudinal data systems

Evaluation of synthetic data

- Synthetic data research utility assessment
  - *Do you get the “right” answer from the synthetic data?*

- Disclosure risk assessment
  - *Do the synthetic data pose a risk of disclosure?*

Scope of GSDS

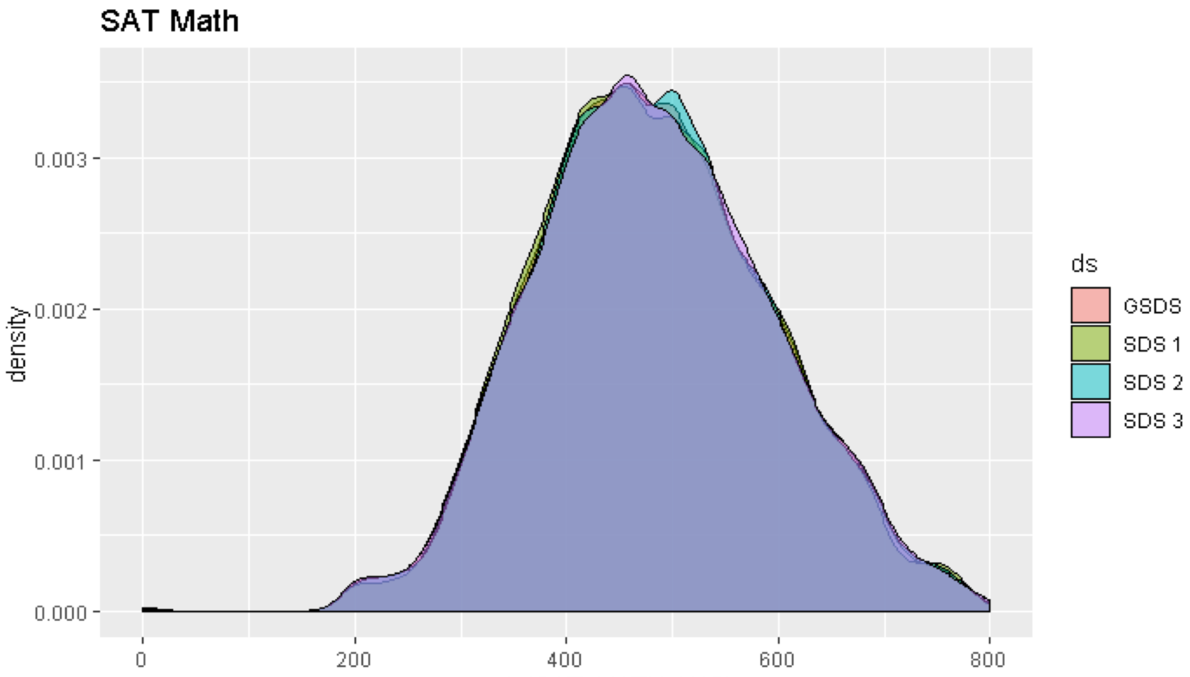
- GSDS is comprised of data from:
  - High school students that entered the workforce
  - High school students that enrolled in post-secondary programs
  - Post-secondary students that entered the workforce

- In total, ~ 100 unique variables in the GSDS
  - Measures for many aspects of education in high school and post-secondary programs
  - Repeated measures for individuals on many variables over time (e.g., GPA, wages)

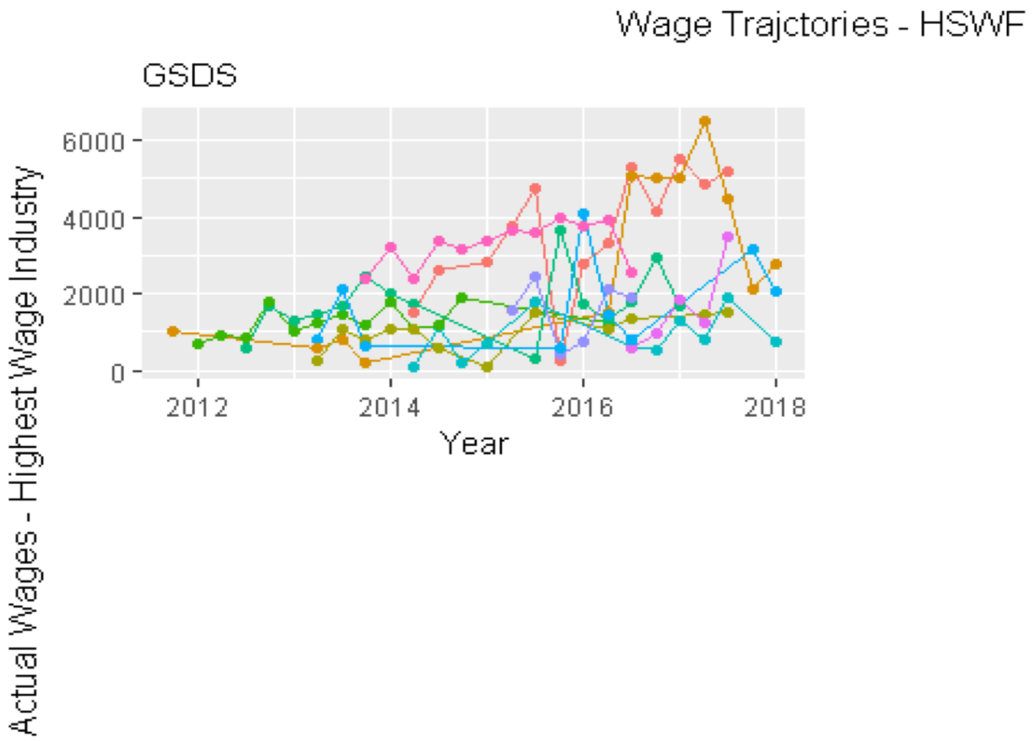
MLDS Center
Maryland Longitudinal Data System

Utility Assessment

- Comparisons of variable distributions
  - Histograms and density plots



Utility Assessment

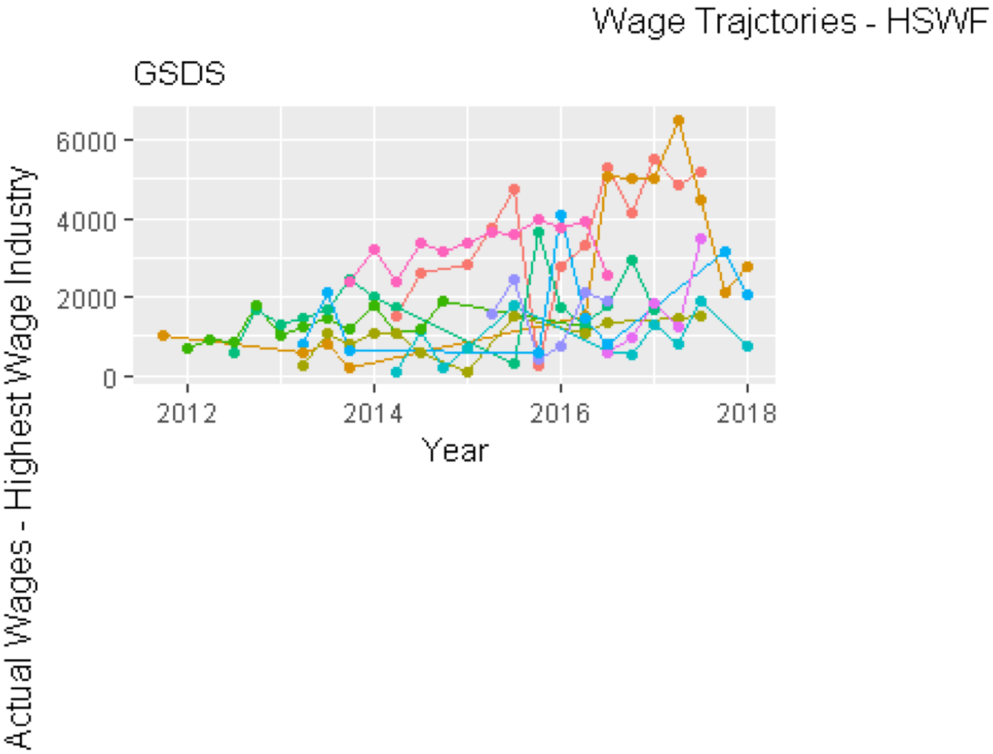


Utility Assessment

- CART model was not well tuned for wages
- Only one lag was used for employment in each sector
- Quarterly wage by sector was creating sparse data

- The solution that was implemented is the following:
    - All possible lags for wages are now used in the predictor set
    - Yearly global wage is synthesized first with all lags
    - then quarterly percentages with all lags
    - then sector percentage within quarterly with same sector lags and all quarters

Utility Assessment



Year

Utility Assessment

- Comparisons of descriptive statistics
  - Means and standard deviations
  - Ranges for continuous, factor levels for categorical
  - Proportions of missing values
  - Correlations, contingency tables

- Evaluate within subgroups (e.g., Male/Female)

Utility Assessment - Specific

- How well does synthetic data reproduce the results of specific analyses?

- Gold standard analyses
  - Standardized mean differences
  - Bivariate correlations
  - Multiple regression
  - Logistic regression
  - Time series

Utility Assessment - Specific

- To illustrate components of utility assessment, we use a subset of the HS->WF GSDS and three SDSs.
- Regressed (log transformed) 2016 wages on gender, SAT-Math, transformed 2015 wages, and race/ethnicity categories
- The sample size of this cohort was 51,863 students
- We calculate the standardized difference between the estimates of interest based on the GSDS and for each SDS as

$$SD = \frac{\beta_{SDS} - \beta_{GSDS}}{SE_{GSDS}}$$

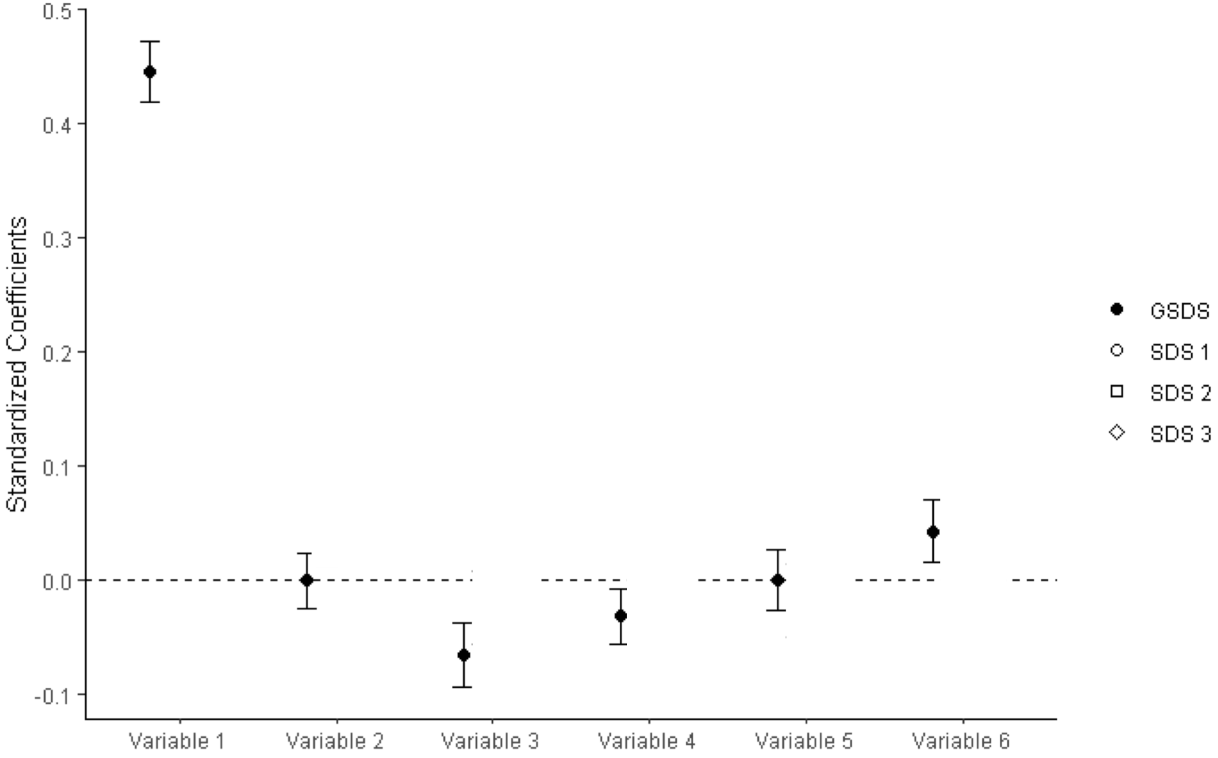
Utility Assessment - Specific

- We also calculate the measure of confidence interval overlap for each estimate (Karr, Kohnen, Organian, Reiter, & Sanil, 2006) as

$$IO = .5\left\{\frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{GSDS} - LCL_{GSDS}} + \frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{SDS} - LCL_{SDS}}\right\}$$

- where UCL_{SDS} and LCL_{SDS} represent, respectively, the average upper and lower confidence limits for the replicated estimates based on the SDSs and where UCL_{GSDS} and LCL_{GSDS} are the confidence limits for the estimate based on the GSDS
- Note that when the two confidence intervals do not overlap, the further they are away from each other the more negative the *IO* estimate will become.

Utility Assessment - Specific



Utility Assessment - Specific

Predictors	GSDS *B* (SE)	AVG SDS *B* (SE)	SD	CI Overlap
Variable 1	0.446 (0.014)	0.343 (0.033)	7.572	-0.152
Variable 2	0.001 (0.012)	0.047 (0.014)	3.823	0.107
Variable 3	-0.065 (0.014)	-0.001 (0.018)	4.526	-0.018
Variable 4	-0.031 (0.012)	-0.007 (0.015)	1.912	0.568
Variable 5	0.001(0.014)	-0.004 (0.015)	0.358	0.914
Variable 6	0.043 (0.014)	0.01 (0.016)	2.365	0.443

Utility Assessment - Specific

Predictors	GSDS *B* (SE)	AVG SDS *B* (SE)	SD	CI Overlap
Variable 1	0.446 (0.014)	0.343 (0.033)	7.572	-0.152
Variable 2	0.001 (0.012)	0.047 (0.014)	3.823	0.107
Variable 3	-0.065 (0.014)	-0.001 (0.018)	4.526	-0.018
Variable 4	-0.031 (0.012)	-0.007 (0.015)	1.912	0.568
Variable 5	**0.001(0.014)**	**-0.004 (0.015)**	**0.358**	**0.914**
Variable 6	0.043 (0.014)	0.01 (0.016)	2.365	0.443

Utility Assessment - General

- How well does the synthetic data reproduce the variable relationships in the GSDS
    - Not tied to a specific analysis

- Several methods have been proposed
    - Kullback-Leibler divergence
    - Cluster analysis
    - Propensity scores

Utility Assessment - General

- Propensity score method

	Dataset	Variable 1	Variable 2	Variable 3
Real Data	0	0	4	1
	0	0	6	1
	0	1	9	3
	0	1	12	5
	…	…	…	…
Synthetic Data	1	0	5	0
	1	0	5	0
	1	1	8	0
	1	1	12	0

Utility Assessment - General

- Propensity score estimation
- Logistic regression
  - Interaction terms for higher-order moments
  - Generalized additive model

- Statistical learning
  - CART
  - Random forest
  - Boosted trees

Utility Assessment - General

- Overall measure of utility (Snoke, 2018; Woo, 2009)
  - Mean square error of propensity scores (pMSE)
    - pMSE → 0, less discrepancy between real and synthetic datasets
  - Mostly used for comparing data synthesis methods

- Variable importance
  - Variables with high importance indicate discrepancies between the GSDS and SDS

Disclosure Risk Assessment

- Identification disclosure
  - *relates to the potential for an intruder to match a given record with a specific individual*
- Attribute disclosure
  - *refers to the possibility that even aggregate data collected from these systems have the potential to disclose aspects of different subpopulations that may be sensitive in nature*

Assessing Risk: Identification Disclosure

- Identification Disclosure rests on the assumption that the synthesized data contains identifiable information about individuals from the GSDS on which it was modeled

- For fully synthesized data the “cases” do not exist (there are no “real” records), so theoretically, there is no identification disclosure risk (the probability would conservatively be 1/N)

Assessing Risk: Attribute Disclosure

- Attribute Disclosure relies on utilizing outside information (such as an additional dataset) to create inferences as a means to identify at-risk groups (<10)
- To assess the attribute disclosure risk we are using a subset of the original GSDS as our "outside source" of information
- The use of the original data provides a worst case scenario of external information an intruder might possess
- Disclosure risk is calculated as the odds of determining sensitive information (such as wages or test scores) using a process of probability matching between the synthetic and "outside" data

Summary

- Public release of synthetic data has the potential to substantially expand access to the MLDS

- Research utility
  - Multiple methods of assessment
  - Results inform data synthesis model

- Disclosure risk
  - Identification and attribute disclosure
  - Because all variables are synthesized, in general disclosure risk is low

Thank you!

- Contributors:
  - Daniel Bonnery, Yi Feng, Angie Henneberger, Tessa Johnson, Mark Lachowicz, Bess Rose, Terry Shaw, Laura Stapleton, Mike Woolley
  - Email: mlachowi@umd.edu

- Acknowledgement:
  - This presentation was prepared by the Research Branch of the Maryland Longitudinal Data System Center (MLDSC) as part of funding from the U.S. Department of Education (R372A150045)
  - The Research Branch would like to thank the entire staff of the MLDSC for their assistance with the work and the presentation.